to focus on this concept. Indeed, the entire section on Inferences seems misplaced. Should it not be at the end of the document? On the other hand, Figure 1 is useful to the discussion of representativeness. The branches in which one must assess this factor offer an excellent opportunity to introduce techniques, etc., to assess representativeness. For example, the figure instructs the reader to follow the algorithms outlined in checklists I-IV. Why not discuss them now? It would seem that a discussion of Figure 1 in light of representativeness would be a more useful first step than to develop concepts of inference form it. The figure is designed to result in an inference, granted, but the pedagogical role of the figure here is to help the reader understand the concept of representativeness.

The next section, **Components of Representativeness,** begins to dissect the concept into pieces more manageable. The table, Table 1, and the coupling of the discussion to the Checklists in the appendix, are perhaps the strongest parts of the Issue Paper. Table 1 is especially noteworthy. It presents the fundamental questions and parses them out according to the “population” characteristics under investigation.

These include Individual Characteristics, Spatial (here misspelled as “Spacial”) Characteristics, and Temporal characteristics. Further, the characteristics are divided between exogenous and endogenous effects- a very useful division. The focus should remain on this table. Discussion should expand, examples given, and understanding reached. These are the essential concept of the Issue Paper.

Unfortunately, the manuscript gets bogged down a bit at this point with the “Case” scenarios. I kept getting confused between Case 2, Case 2a, etc. Also, the introduction of the National Food Consumption Survey confused rather than helped. I found myself wondering if this approach was only applicable to the NFCS or did it have more general applicability. The topic is very general and the specificity of the example obscured that. Again, the tabular presentation is much more straightforward and helpful. Table 2 could be discussed without reference to the NFCS and the different components of representativeness addressed much more clearly and generally.

With section 5, **Attempting to Improve Representativeness,** the tenor of the Issue Paper changes dramatically to become much more statistical in nature. It also becomes more difficult to follow. At points in this section, the authors go off on tangents. See for example the discussion on raking techniques on page 12. A better approach would include more on when such data are likely to be suspect and a better description of the weighting techniques that have been advocated.

In the sub-section **Adjustments to Account for Time-Unit** Differences, there is considerable discussion of the Wallace, et al., approach to inferring temporal effects. No mention is made, however, of the work of Slob (See Risk Analysis 16, 195-200, 1996) who advocates a different technique and evaluates both. Regardless of this missing reference, one questions why it is here at all. It is very detailed and, in my opinion, should be described briefly in terms of its logic, then detailed in an Appendix. The brief reviews of the Clayton et al., paper, the two Buck, et al., papers, the work by Carriquay and co-workers should receive the same treatment.

The section **Summary and Conclusions,** is really only a summary. The first two paragraphs perhaps should have come earlier in the document rather than at the very end. They express the philosophy of what needs to be done. This is a good thing- it sets the stage for the Issue Paper.

Continued Thoughts

After the above impressions while reading the document, I have come away with the impression of a fairly uneven presentation that may not be especially valuable either to the risk assessment community nor to EPA. The idea of an Issue Paper addressing the concept of representativeness is a good one. Data are often used in a willy-nilly fashion with little regard for the way in which they were collected not what the study design intended to do. Because of this, erroneous conclusions can be drawn resulting in much wasted effort and, sometimes, money.

I think the document as now presented does not present the issues well. However, the Figure, Tables, and Checklists are excellent. They provide a strong foundation for a document useful for both the neophyte and expert alike. As an exposure assessor, I am always trying to come up with clean definitions of the parameters, I am measuring. IS it exposure? Is it dose" Is it applied dose? The authors of this Issue Paper draft have crafted answers to similar questions associated with the representativeness of data, surrogates for data, and the pitfalls of ignoring the problem altogether. Unfortunately, these gems are buried in a veritable rockslide of other information. They are not given their proper attention in the Issue Paper. The science and EPA would be well served by asking for a re-write based on the Figure, Tables, and Checklists. Some introductory prose should be placed up front to set the stage- perhaps the two paragraphs (or modifications thereof) found at the beginning of the Summary. This material would be descriptive of the problem at hand answering questions such as why

is representativeness critical, how is it often lacking, and why attempts to improve the representativeness of sample must be done carefully. This would then be followed by Figure 1 and its description, which leads further on to Table 1. The description of Table 1 and Figure 1 give the essentials of the representativeness argument.

The next section would use Table 2 as its focus. Table 2 expands on the ideas of Table 1 and thus is an excellent follow on. The “examples” could be relegated to an appendix with more complete examples chosen and more detailed calculations worked out.

Finally, the Checklists should be given a more prominent placement, and a more complete discussion.

Mitchell Small

**Comments on Pre Workshop Issue Papers:**

**"Evaluating Representativeness of Exposure Factors Data," and "Empirical Distribution Functions and Non-parametric Simulation" for US EPA Workshop on Selecting Input Distributions for Probabilistic Assessment**

**(New York, NY; April 21-22, 1998)**

## Issue Paper on Representativeness

### Overall

I find the discussion of representativeness in this first issue paper to be generally thoughtful and helpful. The paper does a good job of presenting statistical concepts of experimental design in a manner that should be understandable to most exposure modelers. The major issues of target populations versus sampled or surrogate populations, and differences in available vs. desired spatial and/or temporal coverage and scale, are addressed in a clear and comprehensive manner.

### *Tiered Approach* and *Sensitivity Analysis*

The issue of tailoring the framework to a tiered approach to risk assessment is integrally linked to the importance and need for sensitivity analysis when the tiered approach is used. When simpler screening level assessments are pursued, sensitivity analysis is critical to determine whether a significant problem, worthy of attention or remediation, could occur. Sensitivity analysis is always most meaningful in a decision analytic framework - can the decision derived from the risk assessment change as a result of a change in the simplifying assumption (in this case, the use of data or distributions derived from a sample of questionable representativeness)? The only way to determine whether this is so is to repeat the analysis with the underlying data or derived distributions modified in a manner consistent with known or suspected differences, over the range of plausible adjustments.

If a plausible adjustment does lead to a change in the risk management decision, then the analyst must first consider a more rigorous basis for determining the adjustment. If, with a better basis for making the adjustment, the range of predicted exposure or risk still "straddles" multiple decisions regimes (i.e., different management decisions are still possible given the improved adjustment and the overall uncertainty from other assumptions/parameters in the assessment), then this suggests the need to move to the next level of sophistication in the tiered approach. This could include the use of a more detailed and rigorous exposure and risk assessment model, as well as collection of a more representative sample for the target population.

*Adjustment*

The discussion of methods for modifying statistical estimates derived from a surrogate population to obtain results applicable to a different target population is thorough and informative. I do have a few insights to add on encouraging the use of hierarchical models with covariates to derive more representative distributions for the target population; on variance adjustment methods for spatial data; and on the use of Bayesian methods for combining information from surrogate (e.g., national) and target (e.g., site-specific) samples.

Adjustments based on covariates: The discussion in Section 5.1 covers the usual methods for weighting sample observations or sample statistics to adjust for stratification of the target population in the sampled population (either intended, as is the case in a pre-planned survey of the target population, or unintended, as is case addressed in the issue paper, when the stratification weights are a matter of happenstance). The discussion does recognize the utility of covariates (either continuous or discrete) for determining sample weights and mentions the method of "raking" for deriving these.

I think more could be done to encourage the collection and use of covariate data, in particular, using these data to develop "derived distributions" for the target population. Derived distributions arise when a relationship between the parameter of interest and the covariates can be established in a surrogate population. [This relationship could be modified for the target population based on a small sample and Bayesian methods (see my discussion below for how this might be done).] The relationship is combined with the distribution of the covariates in the target population to derive the distribution of the parameter of interest in the target population. The relationship need not be deterministic -the method is quite amenable to use with the usual regression relationships (with explicit distributions of residuals) that are developed in exposure assessment.

Consider the following example with a simple, closed-form solution: For subgroup j (i.e, based on gender, ethnicity, urban vs rural, etc.), the natural logarithm of house-dust lead, ln (house-dust lead), for person k is related to income, I, with the following relationship:

$$\ln (HDL_{k,j}) = a_j + bj[\ln(I_{k,,j})] + e_{k,i}$$

where $a_j$ is the intercept, bj the slope and $e_{k,j}$ the residual of the regression relationship, with

$$e_{k,j} \sim N(0,\sigma_{ej}).$$

If income I for subgroup j is lognormal:

$$\ln(I_{k,j}) \sim N(\xi_{ij}, \phi_{ij})$$

then HDL for subgroup j is also lognormal with

$$\ln(HDL_k) \sim N(\xi_{HDLj} = a_j + b_j\xi_{ij}\ ,\ \phi_{HDLj} = [b_j^2\phi_{ij}^2 + \sigma_{ej}^2]^{0.5})$$

The distribution of HDL for the entire target population with subgroup proportions $P_j$, is the $P_j$-weighted mixture of the lognormal distributions determined for each subgroup.

For more complicated relationships between the parameter of interest and the covariates, or a more complicated distribution of covariates in the target population, Monte Carlo simulation methods may be required to derive the distribution. An example of this (entitled, "Bayesian Analysis of Variability and Uncertainty of Arsenic Concentrations in U.S. Public Water Supplies," by Lockwood, et. al.) is attached. It presents early results of a project for the EPA Office of Ground Water and Drinking Water (OGWDW) to estimate a national distribution of arsenic occurrence in source water used by drinking water utilities, based on a stratified national survey. The application is an example of Case 3 in Table 2, where the surrogate population is a subset of the population of concern. The most pertinent part of the attachment is highlighted, noting that the national distribution is synthesized by sampling the covariates of the target population.

The use of covariates for deriving distributions of exposure factors in a target population is a powerful tool that should be encouraged in the issues paper with more examples and methods. It would also encourage exposure assessors and analysts to be more careful and thorough in their collection of covariate data as part of their monitoring programs.

Variance adjustment for spatial data: The report does a good job covering the options for adjusting bias and variance for time-unit differences; similar methods can be utilized for differing scales of spatial representation. A good reference for this is Random Functions and Hydrology (Bras, R.L. and I. Rodriguez-lturbe, Addison Wesley, Reading, PA, 1985) especially Section 6.8, Sampling of Hydrologic Random Fields. Methods are presented for accounting for spatial correlation when determining the variance of an area average. (The other thing we should do is vote on the correct spelling of spatial/spatial.) Bayesian methods for combining information from surrogate- and target-population samples: I have learned a lot recently about Bayesian methods for combining expert judgment and observed data to estimate distributions. Some of these are discussed in the attached

paper by Lockwood et al. The Bayesian method allows a prior judgment for distribution parameters to be updated based on an observed data set, yielding a posterior distribution for the distribution parameters. The posterior distribution characterizes the uncertainty in the resulting estimation, but can also be used for “best-fit,, point estimates (e.g, based on the mean or mode of the posterior distribution). Bayesian estimates converge to those of classical methods when “vague” or “informationless” priors are used, so that the information in the sample dominates that of the prior.

Bayesian methods can add a lot to the suite of tools available for using surrogate population samples when estimating target population statistics. A number of these tools are described in a paper that Lara Wolfson and I are (hopefully!) about to complete, “Methods for Characterizing Variability and Uncertainty: Bayesian Approaches and Insights” (we have been “about to finish this paper,, for quite a long time, covering a few of our recent meetings - hopefully I will bring a copy to the meeting in New York). In particular, estimates from surrogate population samples can serve as priors for the target population, allowing information from (presumably small and limited) site-specific studies to be informed by, and combined with, the previous studies of the surrogate population. Results from multiple surrogate populations can also be used, each given a weight, along with the informationless prior, to determine how much the resulting estimate will be based on each of the surrogate population studies vs. the information in the target population survey itself.

**A Specific Comment on the Representativeness Paper:**

The discussion on “Summary Statistics Available” in Section 5.1 (page IO) contains what I believe to be an error, when suggesting that standard deviations be averaged across subgroups when approximating a population standard deviation: “In the case of population variance, we recommend calculating the weighted average of the group standard

deviations, rather than their variances, and then squaring the estimated population of concern standard deviation to get the estimated population of concern variance." However, neither of these approaches properly accounts for possible differences in the means across the subgroups, which also contribute to the population variance. The correct approach is to compute E[X2] for each subgroup:

$$E[X_g^2] = E^2[X_g] + Var[X_g]$$

then $E[X^2]$ for the population:

$$E[X_{ATP}^2] = \Sigma_g P_g E[X_g^2]$$

and finally, the variance of X for the population:

$$Var[X_{ATP}] = E[X_{ATp}^2] - E^2[X_{ATP}]$$

where $E[X_{ATP}]$ is computed using the middle equation on page 10.

**Issue Paper on Empirical Distribution Functions and Non-parametric Simulation**

You appear to have already gathered a lot of thoughtful comments on the two topics addressed in this issue paper. Will any of these respondents be at our meeting? Will they be identified? I have given more thought to Part II (Issues related to fitting theoretical distributions) than I have to Part I (Empirical distribution functions). I identified strongly with the comments of Respondent #6 in Part II. To add slightly to Respondent 6's comments, I note that parametric tests of significance for the fit of a TDF almost always reject a particular parametric form as the sample size gets large-real populations invariably exhibit some deviation from a theoretical model, which cannot capture all of the population's behavior and nuances. In these cases, visual comparisons of observed and fitted distributions are essential for determining whether these deviations are in fact important to the problem at hand.

**Review Comments on "Issue Paper on Evaluating Representativeness of Exposure Factors Data."**

**(March 4, 1998 Report)**

by Edward J. Stanek Ill

Are questions of differences in populations, questions related to differences in spatial coverage and scale, and Questions related to differences in temporal scale complete? Should other areas be added?

The document defines a population in terms of a set of units (subjects) at a location and time, a definition that is a standard starting point for traditional survey sampling. The definition of the population is important, since the term "representativeness" is being used to describe the relationship between estimates of exposure, and the true exposure of subjects in the population (or summary measures of these true exposures). An example of a typical population is (p3) "the population surrounding a Superfund site".

The population is defined as a "snapshot" of persons in time and space. Although this definition fits the traditional survey sampling paradigm, this definition may be lacking from the stand point of defining exposure in the context of the public's health. The photographic like quality of the definition does not account for the fact that new people may move into the picture, and others may leave after a short time has passed. Thus, while "representativeness" may be assessed for the picture, the picture itself may be limited. As a result, the assessment of representativeness may have limited relevance for exposure and ultimately the public's health. Of course, when one looks at the "snapshot" close to the time it was taken, the differences may be slight. After a longer time period, the

differences may be dramatic. This practical concern over defining the “population” is ignored in the report.

It is important to introduce a longer time frame and possible changes in exposure when defining exposure in a population. Such definitions are important conceptually, pragmatically, and politically since they define the target parameters for exposure. Such definitions are accessible to a broad range of interested parties and not limited to statistical or technical experts. They set the stage for decisions on additional data collection, and technical choices for estimation and modeling. The current document limits the scope of “representativeness” by defining it only in a context that has an established traditional statistical literature.

In a simple sense, such a definition may be diagramed as in Table 1. The idea is that over chronologic time, there will be mobility and other physical changes. Thus, exposure for the first subject (ID=1) may differ between 1998 ($E_{,,}$) and 1999 ($E_{,12}$). Similarly, ID=1 may move in the year 2000, and hence no longer be exposed. Other subjects may move in the area. Subjects will also age, and their exposure may change with age. Of course, the exposure values in Table 1, while potentially observable, are not known. Nevertheless, a consensus on what will constitute such a potentially observable exposure table is the starting point for discussion of “representativeness”. This conceptual framework has a rich background (Little and Rubin (1987)).

The present document defines the problem in terms of the shaded cells in Table 1. I suggest that the starting point should more closely correspond to a population as defined in Table 1. Establishing the goal first will help prioritize issues such as representativeness, sensitivity, and adjustments. One might dispute this goal by arguing that the problem definition is difficult, exceedingly complex, and since conceptual, detracts valuable time

and effort from what data is known. I would argue that establishing consensus on this definition (while not statistical) should be the starting point for "evaluating representativeness". The definition itself is likely to force an expansion of the context to more modern sampling literature, such as super-population models and methods and model-based inference (Cassel et. al., (1977) Scott and Smith (1969), Meeden and Ghosh (1997)).

Table 1. Potentially Observable Exposure on Subjects in the Defined Spatial Location

($E_{ij}$ =Exposed )

| Time (Yr) (j) | Subject IDs (I) | | | | | | | | Average |
|---|---|---|---|---|---|---|---|---|---|
| | ID=1 | ID=2 | ID=3 | ... | ID=N | ID=N+1 | ID=N+2 | ID=N+3 | |
| 1998 | $E_{11}$ | $E_{21}$ | $E_{31}$ | ... | $E_{N1}$ | | | | $\mu_{1998}$ |
| 1999 | $E_{12}$ | | $E_{32}$ | ... | $E_{N2}$ | $E_{N+1,2}$ | | | $\mu_{1999}$ |
| 2000 | | | $E_{33}$ | ... | $E_{N3}$ | | $E_{N+2,3}$ | $E_{N+3,3}$ | $\mu_{2000}$ |
| ... | | | | | | | | | |
| | $\mu_1$ | $\mu_2$ | $\mu_3$ | ... | $\mu_N$ | $\mu_{N+1}$ | $\mu_{N+2}$ | $\mu_{N+3}$ | $\mu$ |

Are there ways of formulating questions that will allow a tiered approach to risk assessment (a progression from simpler screening level assessments to more complex assessments)?

A general strategy for tiering estimation approaches is by ordering the assumptions. With very extensive assumptions, all exposure assessments are easy. For example, assume that everyone at every time in every location has the exact same exposure, and that this exposure can be measured without error. Using these assumptions, a single measure on

a single subject will suffice. These assumptions are clearly too strong to be broadly acceptable, Nevertheless, these assumptions represent an extreme which has as an opposite extreme the target “potentially observable” population (which is exceedingly complex). A gradation of assumptions can be formed between the two extremes, with such a framework leading to a tiered approach.

The framework asks how important are (or sensitive is the analysis to) population, spatial, and temporal differences between the sample (forwhich you have data) and the population of interest. What guidance can be provided to help answer these questions?

The document addresses the way the “surrogate population” represents the population, how the sample from the surrogate population relates to the surrogate population, and finally, how the measured value relates to the true value for the measured unit, Assuming that the population defined is the potentially observable population of interest, this is a good framework for developing inference. Some guidance can be provided to structurally evaluate the sensitivity of the exposure estimates to analysis decisions. To do so, we build estimates from the data to the surrogate population, and finally to the population.

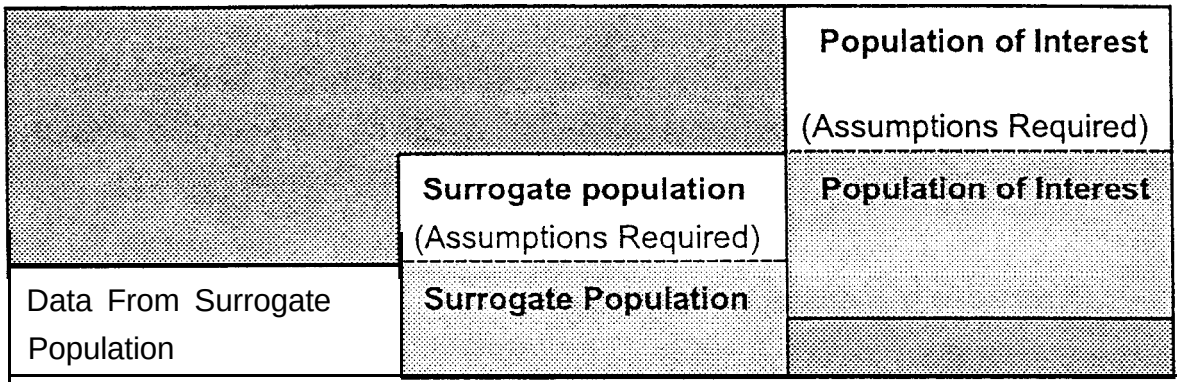
Table 2 represents a framework for successive development of estimates to the population. Probability sampling will connect the surrogate data to the surrogate population, and may serve as the basis for inference to the lower shaded portion of the Surrogate Population. Specifically, the inference consists of estimates of population parameters, and the accuracy (mean squared error) of those estimates. Non-response, limited coverage, etc. may require additional assumptions before inference can be extended to the entire Surrogate Population.

Improvements in the accuracy of estimates for the surrogate population may be possible via modeling and/or post-stratification. The models developed on surrogate data may provide support and serve as a structure for assumptions needed to predict exposure in the surrogate population not stemming from the probability sample. For example, models based on surrogate data may develop a strong dependency of exposure on age and gender, but a weak to null relationship with urban /rural geographic location in one state. Assumptions to estimate exposure in another state (the portion of the Surrogate Population requiring assumptions) may be supported by evidence from the surrogate data, although not directly linked by the probability sampling inferential framework. The range of sensitivity analysis (for example, varying the urban/rural exposure relationship) can be established making used of model based estimates when extending inference to the non-sampled surrogate population.

Models and assumptions most likely will be the primary source to generate estimates from the surrogate population to the population of interest. As the distance increases from the actual data, the role of the models and assumptions will increase. This increased role will result in the estimates being more sensitive to the assumptions. Much progress is currently being made in studying issues of sensitivity similar to these issues in

epidemiology, where a similar situation occurs in observational epidemiologic studies (see recent presentations by Wasserman, Rotnitzky et al. (I 998)). Three-dimensional sensitivity plots, such as those developed by Rotnitzky et al., provide a way of visually communicating and identifying the relative importance of assumptions .

Table 2. Conceptual Steps in Developing Inference from Data to the Surrogate Population to the Population of Interest.




Adjustments

The description of adjustments focus on adjustments due to time unit differences. There are empirical ways of dampening short time variation when estimating longer time interval distributions that do not require parametric assumptions (such as the log-normal assumptions illustrated by Wallace et al (1994). Such methods (such as empirical Bayes methods) require some assumptions, but the assumptions may be minimal and subject to verification. More research is clearly needed in these areas. This is however an active research area that is close to providing answers to practical concerns.

References

Cassel, C-M, Samdal, C-E. And Wretman, J.H. (1977). Foundations of inference in survey sampling. John Wiley and Sons, New York.

Little, R.J.A., and Rubin, D.B. (1987). Statistical analysis with missing data. John Wiley and Sons, New York.

Meeden, and Ghosh, M. (1997). Bayesian methods for finite population sampling. Monograph on Statistics and Applied Probability, Chapman and Hall, New York.

Scott, A.J. and Smith, T.M.F. (1969). “Estimation in multistage surveys,” Journal of the American Statistical Association, 64:830-840.

Wasserman, L. (1998). A tutorial on G-estimation. Eastern Regional North American Annual Meeting of the Biometrics Society, Pittsburg, PA., April 1, 1998.

Rotnitzky A. , Robins, J., and Scharfstein, D. (1998). A G-estimation approach for conducting sensitivity analysis to informative drop-out and non-ignorable non-complacence in a randomized follow-up study. Eastern Regional North American Annual Meeting of the Biometrics Society, Pittsburg, PA., April 1, 1998.

Response to Questions on Issues Paper #1 (Representativeness)

Alan Stern, Dr.P.H., DABT
Div. of Science and Research
New Jersey Dept. of Environmental Protection

I believe that the “checklists” are a conceptually sound and thorough guide to approaching the issues of representativeness. The major problem with the issue of representativeness is not what criteria should be evaluated, but what remedies are available. In my experience, the majority of cases where probabilistic analysis is considered in environmental regulation/standard setting involve choosing a generic distribution to represent an essentially unknown population. That is, default distribution assumptions which can be employed in much the same way that standard point estimates are currently employed in (e.g.) the Super-fund Program. Efforts such as the NHANES III project and other data gathering efforts on national and regional scales often provide data of excellent quality foe large scale populations, Notwithstanding that such data are often structured in a way which can permit information on specific subpopulations to be extracted in a representative fashion, we are rarely in a position to know who those subgroups should be in any given instance. While probabilistic analysis holds out the potential for realistic descriptions of the characteristics of real populations and their exposures, it has been, and, I believe, will continue to be rare for specific populations exposed at a given location to be characterized (other than possibly by their geographic location) in a way which will allow appropriate subpopulation data to be extracted from national/regional databases. If such populations were characterized and/or population-specific exposure data were collected in a focused study, then the issue of representativeness would become a more practical consideration. On the other hand, if such focused studies are not done, then there is little or no quantitative basis for considering whether national/regional population data are specific to the given population. Thus, in most cases the external data are likely to

be "disjoint" with respect to the population of concern. In the absence of population-specific characterization (either with respect to demographics, or, preferably with respect to specific exposure), there does not appear to be an objective way of even identifying how the national/regional surrogate data may be biased with respect to the population of concern.

Having acknowledged this practical problem with deriving representative data exposure distributions, I am not sure that, from the standpoint of public health and risk-based regulation, it is necessarily wise that the population of concern be precisely characterized. The reason for this is that precise characterizations of populations are (as recognized in the checklists) precise with respect to individuals, and their location in space and time. Such information is only precise for a specific moment of time. Demographic and land use patterns change over time, and distributional data which are representative for a given population at a point in time may not be representative for the population at the same location several years or decades later. Risk-based regulatory decisions, on the other hand, are intended to be protective of the exposed population into the indefinite future. Too specific a description of a population of concern may, therefore, make a risk-based regulatory decision unprotective of future populations at the given location. Such considerations seem to argue for more generic tailoring of input exposure distributions to include an intentional component of true uncertainty to address the possible, but unknown values which might apply to future populations and land uses. It is not entirely clear how this should be addressed in quantitative terms, but as a starting point, it seems necessary for such generic descriptions to include the range of values which could reasonably be anticipated to apply to a generic population at a site. To the extent that such descriptions are biased with respect to the current population and/or land uses, that bias should (as appropriate) be toward including more of the high risk population than is already present at the site. For example, if the demographic make up of the potentially exposed population at a given site were such

that there were few young children, the generic input distributions should assume that at some future time, the population could have a larger proportion of young children. It may not be necessary to assume that the national or regional demographics shift in a radical fashion (although over time such shifts, do, indeed, occur), but rather to assume that local demographic idiosyncracies are short-lived. Thus, if a specific locality or neighborhood is demographically skewed toward families with older children, or without children, it should be assumed that in the future, the demographics may shift such that the proportion of young children at the local level of a site reflects the overall state, or county proportion. Such assumptions should be based preferentially on analysis of regional population data, and, if such data are not available, on analysis of national data. One obvious problem with such an approach is that adjustments of current local demographics to current regional demographics to account for future local demographic shifts assumes that regional demographic patterns are more stable than local patterns. This may be true in general, but will not necessarily be true in any given instance.

Tiered Approach

The usual rationale for a tiered approach is that it saves the time and effort which would be needed to conduct population and/or site-specific analyses. Computational time per se, however, is not usually a limiting factor in such analyses. Site-specific data collection, on the other hand, is a major undertaking and is generally a limiting factor. Thus, if population-specific data are available and (as above) it is appropriate to base a risk-based regulatory decision on such data, there is no reason to employ a tiered approach to site-specific distributional descriptions. If, as above, regional-specific distributions are more appropriate for risk-based determinations, and such data are available, then, likewise, a tiered approach is not necessary. If, as is usually the case, population site, or region-specific data are not available, and national population-based data are available, such data may be appropriate as the basis for a screening approach. In considering the use of such data in a non-population-specific context,

however, it must be asked to what extent the specific characteristics of the national data might be misleading for screening purposes. Specifically, are the details of the nation distribution in the extreme tails appropriate, even for screening purposes, for a given subpopulation? Given the screening nature of such an assessment, it may be more appropriate to generate and employ generic screening distributions which use quantitative approximations specifically intended for screening such as triangular, uniform and generalized distributions. Such distributions can also be applied when more complete national population distributions are not available. These distributions could describe, for example, relative minimum values, estimated 10% values, most likely values, estimated 90% values and relative maximum values. It is not necessarily clear that such generic distributions would not be more appropriate for screening purposes than national population-based data. Using such generic default screening distributions would have the additional advantage of establishing specific, and easily identified rebuttable presumptions which would form the starting point for site-specific modifications. Thus, starting from a default screening distribution, it might not be necessary to generate a complete site-specific distribution in order to move toward site/regional specificity. Rather, consideration of the default distribution may help focus the need for more specific information, and it might be realized that the most significant difference between the default assumption and the actual site/regional-specific distribution lies (e.g.) in the upper tail of the default distribution. Thus, it might be necessary only to collect data appropriate to modifying the 90% value in the default distribution.